



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 석 사 학 위 논 문

# HisCoM-PAGE: Hierarchical structural component Models for Pathway Analysis for Gene Expression data

계층적 구조 모델을 이용한  
mRNA 발현 자료의 패스웨이 분석

2019년 8월

서울대학교 대학원  
협동과정 생물정보학과  
목 리 디 아

# HisCoM–PAGE: Hierarchical structural Component Models for Pathway Analysis for Gene Expression data

by

Lydia Mok

A thesis  
submitted in fulfillment of the requirement  
for the degree of Master in  
Bioinformatics

Interdisciplinary Program in Bioinformatics  
College of Natural Sciences  
Seoul National University  
August, 2019

# HisCoM-PAGE: Hierarchical structural Component Models for Pathway Analysis for Gene Expression data

지도교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함

2019년 8월

서울대학교 대학원  
협동과정 생물정보학과

목 리 디 아

목리디아의 이학석사 학위논문을 인준함

2019년 8월

위원장	<u>유 연 주</u>	(인)
부위원장	<u>박 태 성</u>	(인)
위 원	<u>이 승 연</u>	(인)

# Abstract

## HisCoM–PAGE: Hierarchical structural Component Models for Pathway Analysis for Gene Expression data

Lydia Mok

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

Although there have been several analyses for identifying cancer–associated pathways, based on gene expression data, most of these are based on single pathway analyses, and thus do not consider correlations between pathways. In this paper, we propose a hierarchical structural component model of pathway analysis for gene expression data (HisCoM–PAGE), which accounts for the hierarchical structure of genes and pathways, as well as the correlations among pathways.

Specifically, HisCoM-PAGE focuses on the survival phenotype and identifies its associated pathways. Moreover, its application to a real biological data analysis of pancreatic cancer data demonstrated that HisCoM-PAGE could successfully identify pathways associated with pancreatic cancer prognosis.

Simulation studies comparing the performance of HisCoM-PAGE with other competing methods such as Gene Set Enrichment Analysis (GSEA), Global Test, and Wald-type Test showed HisCoM-PAGE to have the highest power to detect causal pathways.

**Keyword:** Pathway analysis, Survival phenotype, Hierarchical structural component model

**Student Number:** 2017-29730

# Contents

Abstract .....	i
Contents .....	iii
List of Tables .....	iv
List of Figures .....	v
1      Introduction .....	1
2      Materials .....	6
3      Methodology .....	9
4      Results .....	18
5      Discussions.....	31
Bibliography.....	34
Abstract in Korean.....	40

## List of Tables

<b>Table 1</b> Demographics of study patients.....	7
<b>Table 2</b> Significant pathways for PDAC prognosis using HisCOM- PAGE .....	27
<b>Table 3</b> Comparison method result (Global test) .....	28
<b>Table 4</b> Comparison method result (Adewale) .....	28
<b>Table 5</b> Comparison method result (GSEA) .....	28
<b>Table 6</b> Significant pathway and gene markers in PDAC prognosis .....	30



## List of Figures

<b>Figure 1</b> Kaplan Meier plot of study patients .....	8
<b>Figure 2</b> A schematic diagram of the HisCoM–PAGE model .....	12
<b>Figure 3</b> The empirical type 1 error .....	19
<b>Figure 4</b> Empirical power comparison of 4 scenarios .....	23
<b>Figure 5</b> Two causal pathway power result .....	24
<b>Figure 6</b> Significant pathways identified by various methods .....	26

# Chapter 1.

## Introduction

Over the past several decades, gene expression data analysis has been the most common approach to investigating human diseases, at the RNA level [1, 2]. By analyzing gene expression data, we can gain better understanding of disease etiology and biological mechanisms [3]. Especially for cancer prognosis, genetic information can be more effective in improving prognosis prediction of patients than the prediction which based only on clinical information [4].

Analyzing high-throughput gene expression data, at the pathway level, is effective in two ways. Firstly, grouping thousands of genes by their respective pathways reduces complexity to just several hundred pathways. Secondly, identifying active pathways that differ between two conditions, such as normal and tumor tissues, can have more explanatory power than a simple list of differentially expressed genes (DEGs) [5]. While there is a need for pathway analysis itself,

the pathway method for survival phenotype is rare. Only a few pure methods are proposed for survival phenotype [5–7].

Various cancer prognoses and survival analysis have been consistently reported [8]. For example, pancreatic cancer has a very poor prognosis, compared to other cancers. At the time of diagnosis, less than 20% of pancreatic cancer patients can have surgery, and their postoperative 5-year survival rate is also significantly lower (less than 25%) [9]. Therefore, more accurate pancreatic cancer prognosis, and early detection, are needed.

To build a good prediction model, using gene expression data, for actual clinical application and medical intervention, it is first necessary to identify features related to prognosis. Furthermore, exploring the pathways to which genes belong can provide valuable biological interpretation, and help screen out false-positive genes. In this study, we mainly focus on finding significant pathways that are relevant to the prognosis of pancreatic cancer. Through pathway analysis, our ultimate goal is to identify biological mechanisms that influence the prognosis of disease more clearly.

Since gene set enrichment analysis (GSEA) was proposed, in which the Kolmogorov–Smirnov statistic is used for measuring differentially expressed gene sets, many pathway–

based methods for continuous phenotypes have been developed [10]. For the survival phenotypes, however, there are only a few pathway-based methods available. For example, the global test was proposed for continuous and censored survival time, based on the score statistics from random effects of parameters for association measure [11,12]. Likewise, the Adewale approach was proposed for the survival phenotype by summarizing the association measure from the sum of coefficients from a survival regression model [13]. More recently, the gene set variation analysis (GSVA) method was proposed to handle survival times by estimating the variation of pathway activity over a sample population in an unsupervised way [14]. However, those previous pathway methods are single pathway analyses, so they do not take into account correlations between pathways, and the global test only considers correlations between gene expression values. The Wald test merely sums up the statistics from each gene, to get its pathway statistics, so it does not account for the correlation among pathways. Since some genes may belong to several pathways simultaneously, there is a need for accounting for this nature of genes and pathways.

To account for this issue, we previously developed our Pathway-based approach using Hierarchical structure of

collapsed Rare variant Of High-throughput sequencing data (PHARAOH) method for discovering rare variants by constructing a hierarchical model that consists of collapsed gene-level summaries and entire pathways [15]. PHARAOH is based on the generalized structural component analysis (GSCA) model [16]. Later, we developed our Hierarchical structured component analysis of miRNA-mRNA integration (HiCoM-mimi) method to integrate anti correlated expression of miRNA and mRNA. By extension of PHARAOH, HisCoM-mimi can also account for the biological relationships between a miRNA and target mRNAs [17]. Recently, we developed another extension, HisCoM-GGI for gene-gene interaction analysis, representing a model that not only summarizes common variants into gene levels, but also considers interactions among common variants [18].

In this study, we develop a new pathway-based model for survival phenotypes, based on gene expression data, by taking advantage of our earlier hierarchical model, called “Hierarchical structural Component Model for Pathway analysis for Gene Expression data (HisCoM-PAGE).” As an extension version of HisCoM-mimi, HisCoM-PAGE considers the biological context of gene and pathway hierarchies, in the form of structured components. Using latent variables, a gene

can be collapsed into the structured form, so it can provide significant pathways and genetic markers related to prognosis. Also, HisCoM-PAGE considers the correlation of all pathways, by using a ridge penalty in parameter estimation. HisCoM-PAGE can also successfully examine the effects of individual genes within the pathways.

Through simulation studies, we showed that HisCoM-PAGE performed well, compared to other existing pathway methods for survival phenotype. Application to real microarray data of pancreatic ductal adenocarcinoma (PDAC) patients from Seoul National University showed that HisCoM-PAGE could well identify prognosis-related pathways.

# Chapter 2.

## Materials

### 2.1 Pancreatic ductal adenocarcinoma (PDAC) samples

From 2012 to 2014, 125 PDAC samples were collected by the Department of Hepatobiliary and Pancreatic Surgery of Seoul National University Hospital. All human subjects' studies were approved by the Institutional Review Board of Seoul National University Hospital. In this dataset, mRNA expression levels were generated using Affymetrix (Santa Clara, CA, USA) HuGene 1.0 ST arrays. We selected the mRNA whose expressional variances were ranked in the top 25 percentiles. The clinical information is described in Table1.

**Table1.** Demographics and clinical characters of study patients.

Clinical variables	Variable Description	Descriptive Statistics
Age	Age at diagnosis	63.32(10.064) mean(se)
Sex		Male: 75, Female: 50
Positive LN	Number of cancers transmitted Lymphocytes	(0,1,2) (1st Quan, Median,3rd Quan)
Size	Maximum Tumor Size (cm)	3.574 (mean)
Differentiation	Clinico–pathologic characteristics and prognostic value of various histological types (Well Differentiated (WD), Moderately Differentiated (MD), Poorly Differentiated (PD))	WD: 19, MD: 85, PD: 18, Other: 2 (NA: 1)
Jaundice		Yes: 89, No: 36
7 <sup>th</sup> staging T stage	AJCC staging criteria	1th: 6, 2nd: 3, 3rd: 104, 4th: 12
7 <sup>th</sup> staging N stage	AJCC staging criteria (Number of positive lymph node exist)	Yes: 71, No: 54
Radiation therapy	Radiation therapy after surgery	Yes: 72, No: 53
Chemotherapy	Chemotherapy after surgery	Yes: 94, No: 31
Status	Censoring indicator	Censored: 62, Dead:63
Overall survival time		Median: 25 months



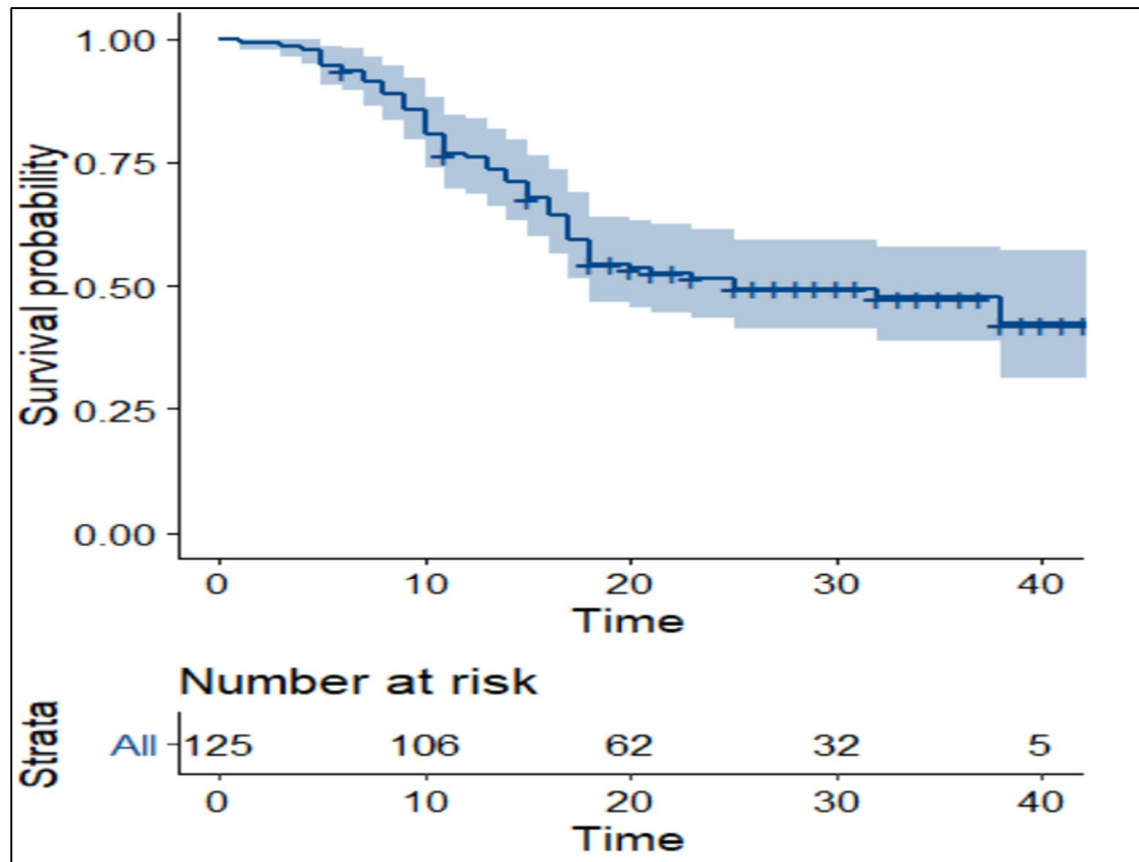


Figure 1 Kaplan Meier plot of study patients.

# Chapter 3.

## Method

### 3.1. HisCoM–PAGE method

Let  $y_i$  denote a survival time ( $i = 1, \dots, I$ ). Let  $x_{jk}$  denote the  $j^{th}$  pathway of the  $k^{th}$  gene expression corresponding to  $j^{th}$  pathway ( $j = 1, \dots, J$ ). As shown in Figure 1, we must then consider latent structures for estimating the model parameters. Let  $w_{jk}$  denote the weight assigned to  $x_{jk}$ . The coefficient  $\beta$  represents the effect of the latent variable  $f_{ij}$  on the phenotype, as  $f_{ij} = \sum_{j=1}^J f_{ij}\beta_j$ . Considering this structure, we designed the following Cox proportional hazard model.

$$\begin{aligned} h(y_i|F_i) &= h_0(y_i) \left( \sum_{j=1}^J \exp\left(\sum_{k=1}^K x_{jk} w_{jk}\right) \beta_j \right) \\ &= h_0(y_i) \exp\left(\sum_{j=1}^J f_{ij} \beta_j\right) \end{aligned}$$

To estimate the model parameters for HisCoM–PAGE, we maximized the penalized partial log likelihood, using a ridge penalty. The following equation then represents the objective function.

$$\begin{aligned}
\phi = & \sum_{i:C_i=1} (\sum_{j=1}^J f_{ij}\beta_j - \log \sum_{l:Y_l \geq Y_i} \exp(\sum_{j=1}^J f_{ij}\beta_j)) \\
& - \frac{1}{2} \lambda_{gene} \sum_{j=1}^J \sum_{k=1}^{G_j} P_{\lambda_{pathway}}(w_{jk}) \\
& - \frac{1}{2} \lambda_{pathway} \sum_{j=0}^J P_{\lambda_{gene}}(\beta_j)
\end{aligned}$$

The objective function can be maximized by an alternating least squares (ALS) algorithm, which iterates the following two steps until convergence. In the first step, the pathway coefficients are estimated and updated, using a least-squares approach. For the second step, the weight coefficients are updated for fixed-path coefficient estimates [16]. In HisCoM-PAGE, we adopted a ridge penalty to address the multi-collinearity of genes within any specific pathway. When estimating  $\lambda_{gene}$  and  $\lambda_{pathway}$  values, we conducted 5-fold cross-validation to obtain optimal values for  $\lambda_{gene}$  and  $\lambda_{pathway}$ . The process of estimating the coefficients, using the ALS algorithm, with penalty, proceeds as follows:

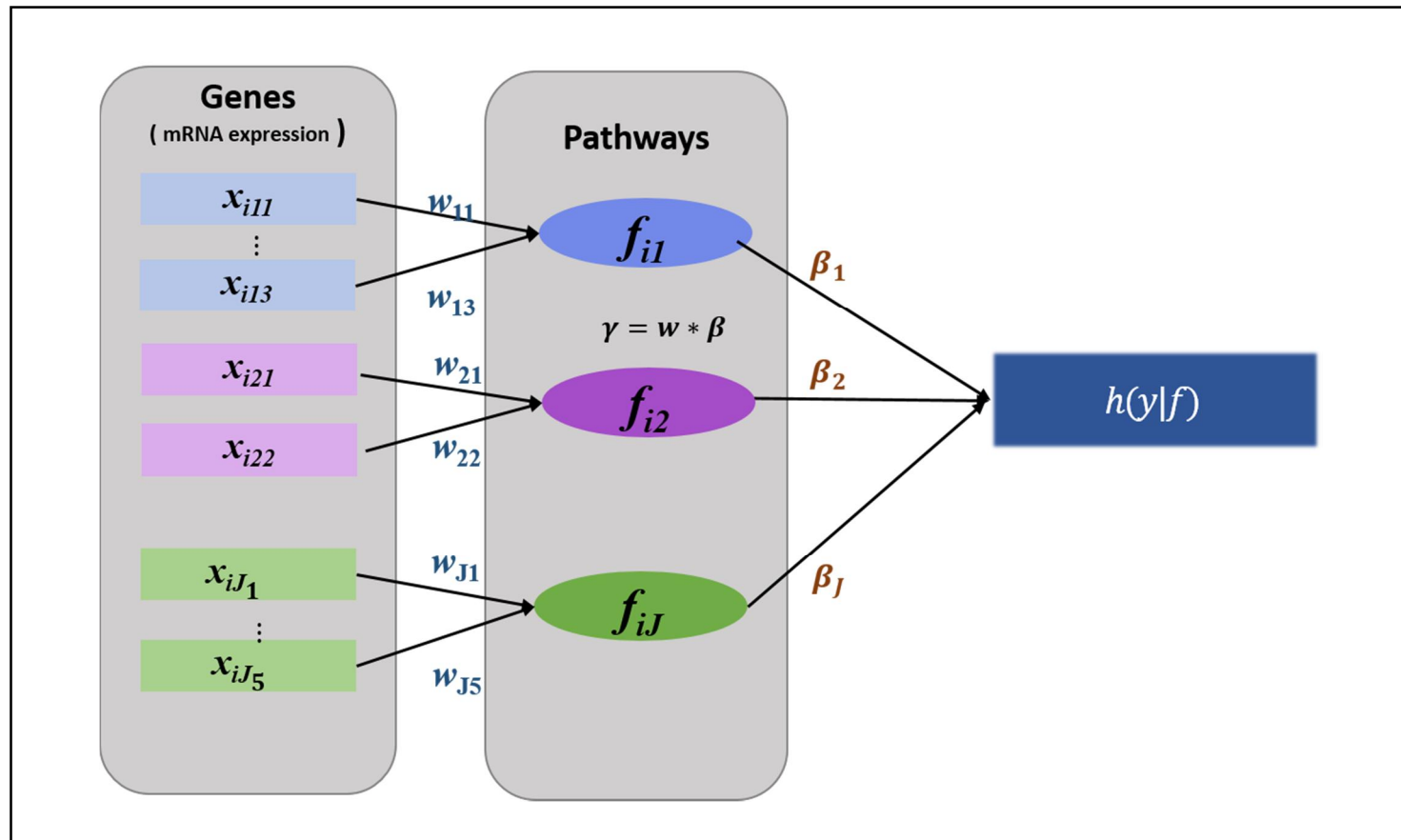
$$\text{Let } \eta = XWB, u = \frac{\partial l}{\partial \eta}, A = \frac{\partial^2 l}{\partial \eta \eta^T}, z = \eta + A^{-1}u,$$

$$\text{then } l(W, B) \approx (z - \eta)^T (z - \eta)$$

1. Fix  $s_w, s_B$ , and initialize  $\hat{B}, \hat{W} = 0$ .

(Where, s represents the residuals.)

2. Compute  $\eta, \mathbf{u}, \mathbf{A}$ , and  $\mathbf{z}$ , based on the latest value of  $\hat{\mathbf{B}}, \hat{\mathbf{W}}$ .  
Then minimize  $(\mathbf{z} - \eta)^T(\mathbf{z} - \eta)$ , with the fixed  $\hat{\mathbf{W}}$  subject, to  $\sum |W_{ij}| \leq s_w$ . Repeat these steps until  $\hat{\mathbf{B}}$  converges.
3. Compute  $\eta, \mathbf{u}, \mathbf{A}$ , and  $\mathbf{z}$ , based on the latest value of  $\hat{\mathbf{B}}, \hat{\mathbf{W}}$ .  
Then, minimize the  $(\mathbf{z} - \eta)^T(\mathbf{z} - \eta)$ , with the updated  $\hat{\mathbf{B}}$  subject, to  $\sum |B_i| \leq s_B$ . Repeat these steps until  $\hat{\mathbf{W}}$  converges.
4. Iterate steps 2 and 3 until  $l(\mathbf{W}, \mathbf{B})$  converges.



**Figure 2.** A schematic diagram of the HisCoM–PAGE model

### 3.2 Comparative methods

The following pathway methods were considered to compare the results of HisCoM-PAGE. We compared other pathway methods such as Gene Set Enrichment Analysis (GSEA) with two type of weight, Global test, and Adewale test with survival phenotype [19].

GSEA methods assume that the total number of genes  $K$  and the gene set  $S$  is predefined. For the first step, compute the regression coefficients of  $k$  genes, by fitting univariate Cox regression models. The regression coefficient is used as an association measure between phenotypes and genes. Secondly, order  $K$  genes by the absolute value of  $t$  statistics ( $= \hat{w} / \hat{se}(\hat{w})$ ) in descending order. Thirdly, calculate the enrichment score. While computing the enrichment score, GSEA methods consider two methods of weighting, including GSEA1, the case when the weight term is 0, and GSEA2, the modified version of the original GSEA method when the weight term is 1 [10,19]. Lastly, calculate the significance level by comparing the observed values and the permutation distribution values.

$$ES = \max |P_{hit}(k) - P_{miss}(k)| = \max |S(k)| \quad (k=1, \dots, K)$$
$$P_{hit}(k) = \sum_{k=1}^K \frac{E(k)}{N_h}, \quad P_{miss}(k) = \sum_{k=1}^K \frac{(1-E(k))}{N_m}$$

The Global test is based on the regression coefficient from a Cox model. Global tests can test whether the expression of gene, within a predefined pathway, tends to closely associate with the survival times. Goeman describes the global test's Q statistic as an average of the m test statistics calculated from each m individual gene, constituting a pathway by itself.

$$Q_j = \frac{T - \hat{E}T}{\sqrt{\widehat{\text{Var}}(T)}}$$

Although the p-values can be calculated using the permutation and asymptotic method, we used the permutation approach [12].

Thirdly, the Wald type test is based on the unified pathway method proposed by Adewale, which combined component-wise test statistics for significance of a subset of genes [13]. Wald tests also assess whether the predefined pathway has an association with survival times.

$$W_j = \sum_{k=1}^K r_k^2 \text{ (where, } r = \hat{w} / \hat{se}(\hat{w}))$$

Thus, the test statistic is a sum of squares of the Wald statistic for the individual genes that constitute to the pathway.

### 3.3. Simulation study

To evaluate the performance of the HisCoM-PAGE method



and compare its performance with other pathway methods, we generated a simulation data set, following the simulation settings of Lee et al [19]. In the simulation study, the following parameters were considered: the sample size ( $I$ ), total number of genes ( $k$ ), pathway size ( $m_s$ ), proportion of censoring ( $c_p$ ), and the proportion of significant genes in the pathway ( $m_p$ ). Gene expression data was randomly generated from a multivariate normal distribution with mean zero and covariance matrix  $\Sigma$ . Four types of different covariance matrices ( $\Sigma$ ) were considered. Let the  $\mathbf{0}$  zero matrix be  $l \times (k - l)$  dimensions, where  $l$  means the number of causal genes within the gene set. Let  $\mathbf{I}_l$  be an  $l \times l$  identity matrix, and  $\mathbf{A}$  be a  $l \times l$  symmetric matrix. Then, the covariance matrix is given as follows:

$$\sum_{k \times k} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0}^T & 0.2\mathbf{I}_{k-l} \end{bmatrix}, (l = m_s \times m_k)$$

For each scenario,  $\mathbf{A}$  has a different structure. Here,  $i, j$  represent each row and column index for covariance matrix. For Scenario 1,  $\mathbf{A} = 0.2\mathbf{I}_l$ ; for Scenario 2,  $\mathbf{A} = 0.2[x_{ij}]$ , and  $x_{ij} = 0.02$ ; for Scenario 3,  $\mathbf{A} = 0.2[x_{ij}]$ , and  $x_{ij} = 0.1^{|i-j|}$ . Scenario 4 has random variances and covariance, such that  $\mathbf{A}$  is given as follows:  $\mathbf{A} = 0.2[x_{ij}]$ ,  $x_{ij} = \rho_{ij}$ , when  $i$  is not equal to  $j$ , and  $1$ , when  $i$  is equal to  $j$ , and  $\rho_{ij}$  is generated from  $N(0, 0.1^2)$ .

The survival time for the subject is defined as the minimum value between the observation and censoring times of the subject. Observation times were generated from the Cox model, in which the baseline hazard function is constant, with a base hazard rate of 0.005. For power analysis, the regression coefficients  $\mathbf{w}$  were generated from the uniform distribution  $\mathcal{U}(0.2, 0.6)$ .

In a simulation setting with one causal pathway, we also considered the following simulation setting for two causal pathways to see how the HisCoM-PAGE would perform. Firstly, we consider the two causal pathways and the two non causal pathways. Secondly, the following parameter also considered in two causal pathways setting: sample size ( $I$ ), total number of genes ( $2k$ ), pathway size ( $m_s$ ), proportion of censoring ( $\mathbf{c}_p$ ), and the proportion of significant genes in the pathway ( $m_k$ ). In the simulation setting of two causal pathways, the gene expression data were modified in the previous one causal pathway setting. For the two causal pathways scenario, covariance matrix ( $\Sigma$ ) be  $2k \times 2k$  matrix.

$$\sum_{2k \times 2k} = \begin{bmatrix} B & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & 0.2I \end{bmatrix} \text{ where, } B = \sum_{k \times k} = \begin{bmatrix} A & 0 \\ 0^T & 0.2I_{k-l} \end{bmatrix}, (l = m_s \times m_k)$$

$A$  matrix is same as the setting in scenario 1 for one causal

pathway. Next, the data were produced assuming that the correlation coefficient between the genes in both causal pathways was equal to 0.5. Lastly, the genes within the same pathway with a correlation coefficient of 0.5, and the genes between different pathways set a correlation coefficient of 0.2.

For type 1 error estimation, the regression coefficients  $\mathbf{w}$  were assumed to be zero. The observation times were generated from the Cox proportional hazard model. The censoring times were generated from an exponential distribution, using the tuning parameter  $\lambda$ , whose values depended on the censoring fraction of the data [20]. Here, the observation times and censoring times were generated independently.

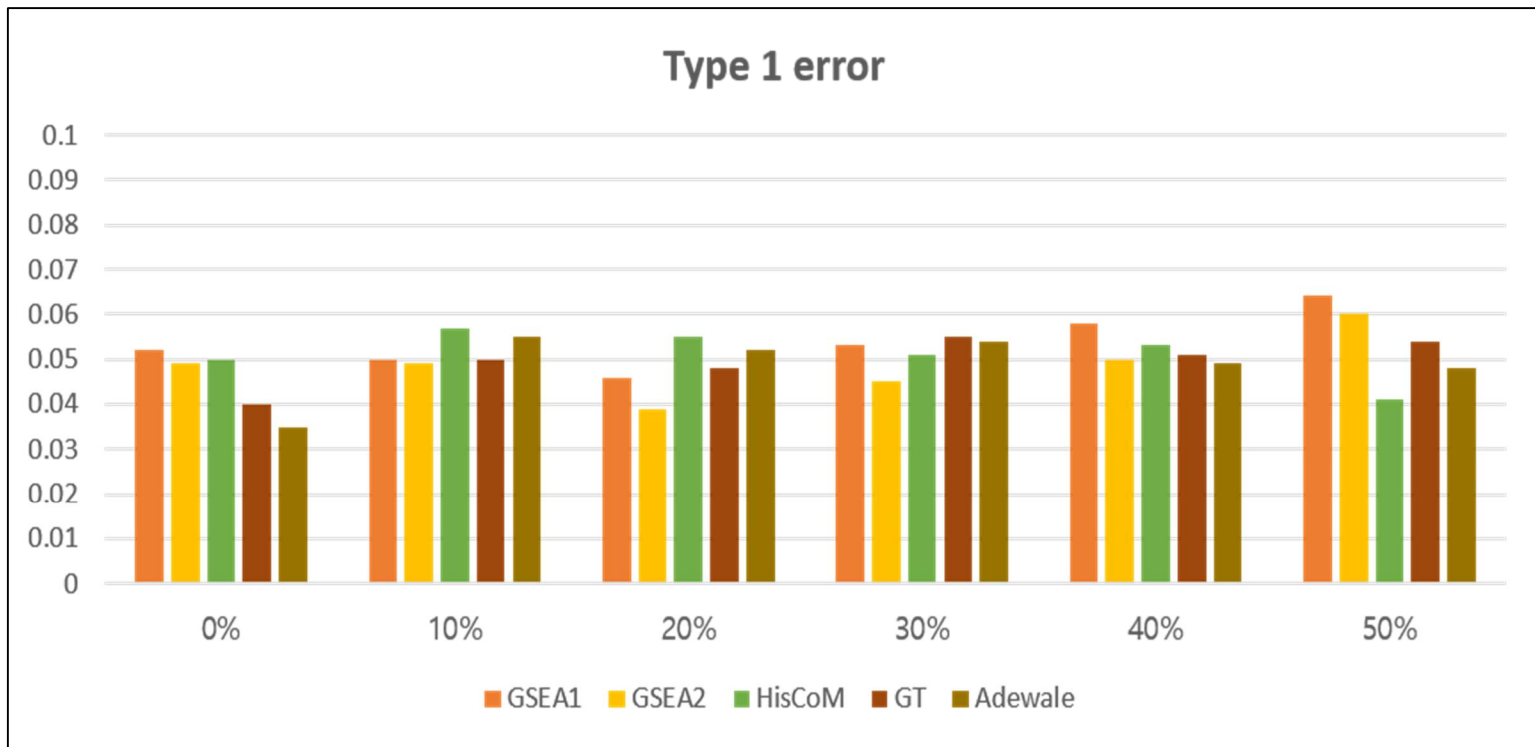
# Chapter 4.

## Result

### 4.1. Simulation analysis result

#### 4.1.1. Type 1 error

Figure 1 shows the simulation results, for each method, when total number of genes ( $k$ ) = 200, sample size ( $I$ ) = 80, gene set size ( $m_s$ ) = 50, and censoring proportion ( $c_p$ ) = 0, 0.1, 0.2, 0.3, 0.4, 0.5. The empirical type 1 error was then estimated with 1000 replicates at a 0.05 significance level. The number of permutations for significance testing was 1000. Overall, type 1 errors were shown to be well-controlled in various scenario settings. Especially in HisCoM-PAGE method, type 1 error is well controlled, even when the censoring fraction is high.



**Figure 3** The empirical type 1 error with 1000 replicates at the various censoring proportion. The x-axis represents the censoring proportion of simulated data.

#### 4.1.2. Power comparison for single causal pathway.

For power analysis, we varied the censoring proportion and the proportion of significant genes in the causal pathway. We set the parameters as follows: total number of genes( $k$ ) = 200, sample size( $I$ )= 80, gene set size( $m_s$ )= 50, and censoring proportion ( $c_p$ )= 0, 0.3. There will be 1 causal pathway and 3 non-causal pathways.

HisCoM-PAGE showed better performance than the other methods, when the significant gene proportions were not high, and the power close to 1, when the significant gene proportion grows larger. Figure 4 shows the power of each method for the four correlation structure scenarios. The x-axis represents a significant gene proportion. Overall, the Global and Adewale methods showed similar trends in power, and GSEA showed a relatively low power, compared to other methods, in many scenarios, as mentioned in Lee' s paper [19]. As shown by Chiristiaan [21], power depends largely on gene proportion, which has effects within a causal pathway. In Scenario 1, i.e., all gene expression values are independent of each other and compared to other scenarios, the statistical power is much affected by the centering ratio. In Scenario 2, the correlation coefficient between casual genes had the same effect, and at this time we could see a

relatively high power, compared to other scenarios.

For the GSEA method, the Cox model was only used to ordering genes, but ES scores were calculated using the relative rank only. In the case of the competitive analysis method of the pathway methodology, observed statistics of the pathway of interest are compared with those of the pathway consisting of the genes within the pathway [22]. By contrast to GSEA-based methods HisCoM-PAGE can directly calculate the effect of the causal pathway, quantitatively, on the survival time, as a Cox model with the structural equation. We could also confirm that the power of HisCoM-PAGE was higher than the other methods in Scenario 1, even when the censoring ratio was high.

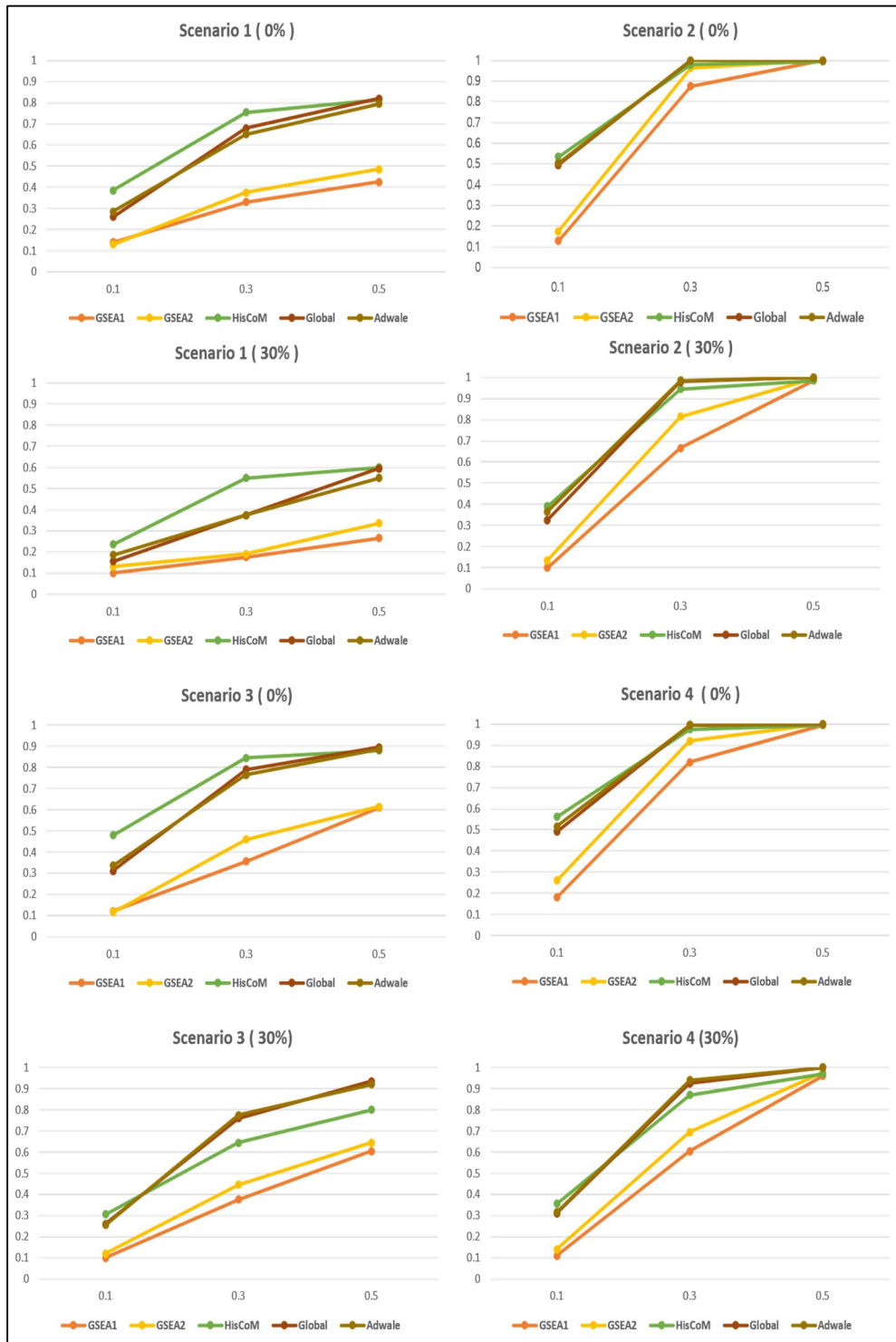
#### 4.1.3. Power for two causal pathways in HisCoM-PAGE.

In addition to the single causal pathway, simulation settings with two pathways having a causal effect enable to identify the following HisCoM-PAGE characteristics.

First, we calculated power from two perspectives: partial power and full power. A partial power is a power for detecting one causal pathway. Full power, which is calculated when both causal pathways were detected. In both partial and full power, they showed a general tendency to increase as the proportion

of causal genes in the pathway increases. When the correlation between genes in a pathway was given the same, and when compared otherwise, full power showed a higher tendency when the between correlation was smaller than the inner correlation. As shown in Figure 5, in the case of two causal simulation settings, which reflect the biological phenomena in which co-occurrence exists between real pathways and co-occurrence occurs between genes, it can be seen that the two causality shows good power in HisCoM-PAGE.





**Figure 4.** Empirical power comparison of 4 scenarios.

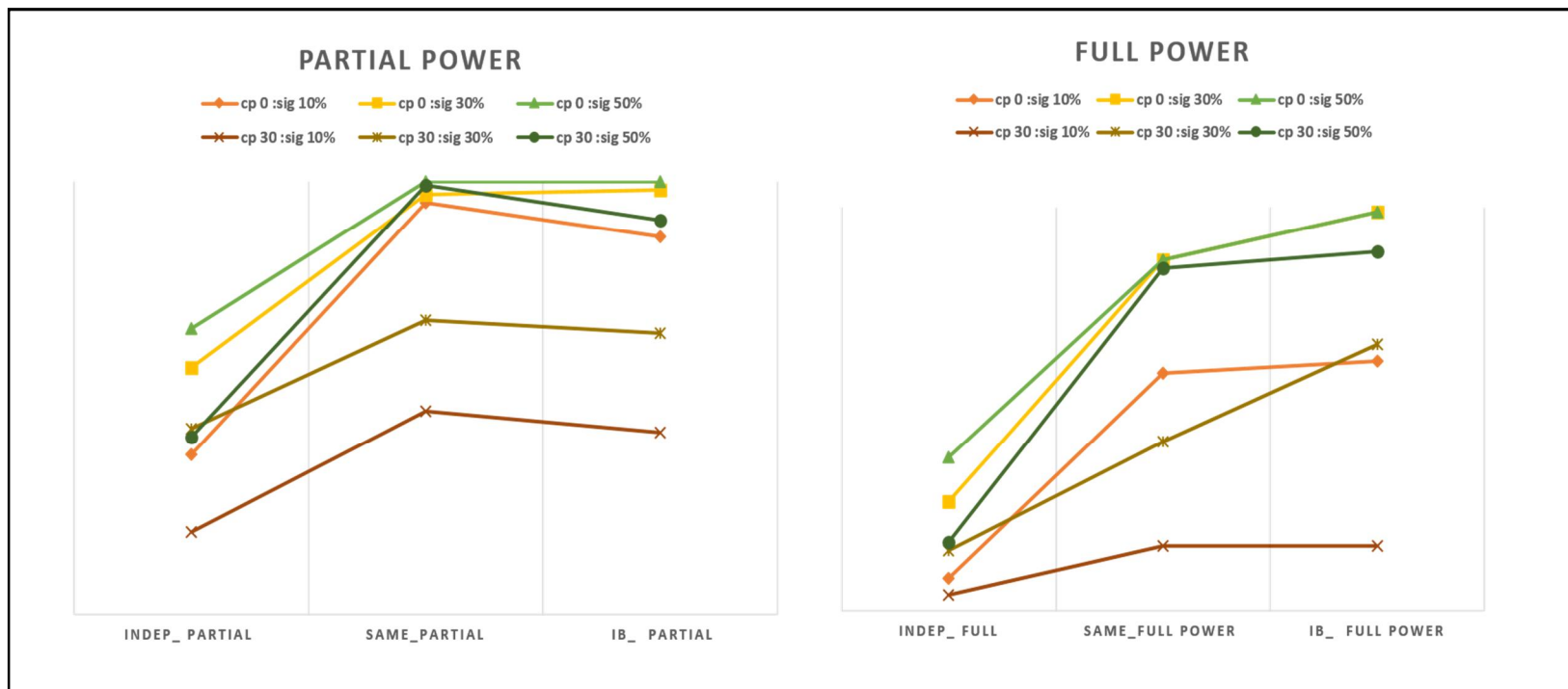


Figure 5 Two causal pathway power result for partial and full power.

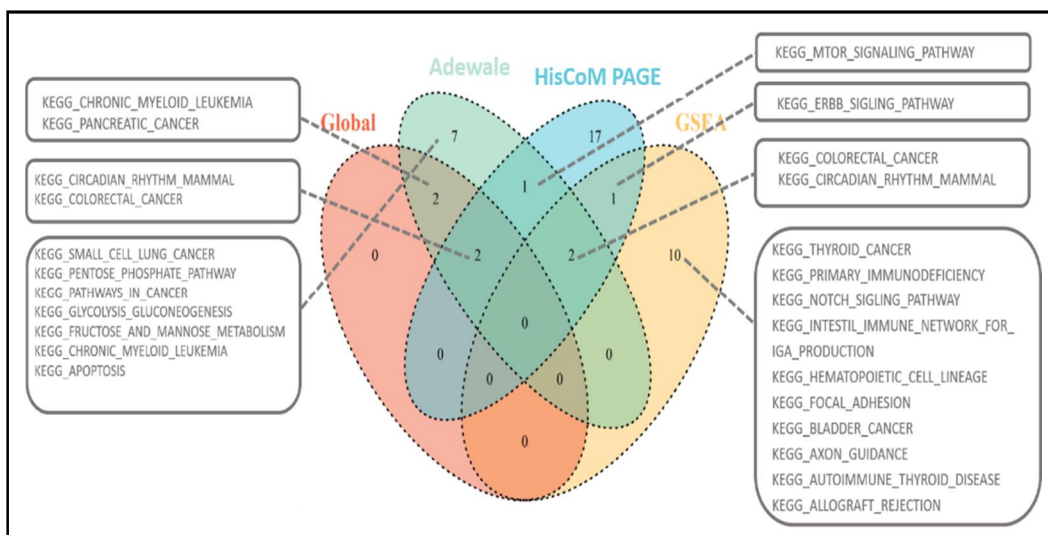
## 4.2. PDAC real data analysis result

### 4.2.1 PDAC related pathways

A total of 32,321 genes were normalized by the robust multi array average (RMA) method [23]. Of the total, 21,369 genes were annotated. In this analysis, we selected mRNAs whose expressional variances were ranked in the top 25 percentiles of the dataset [24–26]. The Affymetrix gene identifiers were mapped to the KEGG and Biocarta databases. Here, we mapped all 4320 genes into pathways. Using the KEGG and Biocarta datasets, HisCoM–PAGE identified 185 pathways and 216 pathways, respectively [27, 28].

Our objective in this pathway analysis was to identify pathways associated with PDAC patient’ s overall survival times. By using the Biocarta database, four pathways were significant after FDR correction, and 23 pathways were significant, based on the KEGG database. Figure 3 shows the top significant pathways related to the survival times from HisCoM–PAGE method. Using the Biocarta database, the TGF–beta pathway was found to be most significant for survival time of PDAC patients. It is well known that TGF–beta pathway associates with inflammation promotion and carcinogenesis in early stage of cancer [29–35]. Figure 4 shows a sampling of pathways commonly identified, although

there are unique pathways found by each method. TOB1, a pathway from the Biocarta database, has previously been reported to be linked to PDAC [36,37], while the Hedgehog pathway was found only by HisCoM –PAGE with regard to PDAC [38]. Hedgehog–signaling dysregulation, due to mutation or overexpression of pathway components and/or pathway ligands, induces pancreatic cancer [38]. The KEGG pathways found uniquely in the HisCoM–PAGE are bolded in Table 2. Figure 4 shows the selected pathways after multiple test correction [39]. KEGG pathways selected significantly in three or more ways are as follows: Adherent junction pathway, Colorectal cancer pathway, Circadian rhythm mammal pathway, and Dorso ventral axis formation pathway.



**Figure 6** Significant pathways identified by various comparison methods.

**Table 2.** Significant pathways for PDAC prognosis using HisCOM-PAGE. A bolded pathway is a representation of a pathways that do not appear significantly in other methods, but only significant in HisCoM-PAGE.

Pathway	Pathway size	$\beta_{path}$	$p$ value	$q$ value
BIOCARTA_TGFB_PATHWAY	8	0.017	0.00001	0.002
BIOCARTA_NTHI_PATHWAY	7	0.014	0.00033	0.030
BIOCARTA_MITOCHONDRIA_PATHWAY	13	0.010	0.00054	0.030
BIOCARTA_TOB1_PATHWAY	14	0.016	0.00056	0.030
KEGG_COLORECTAL_CANCER	19	0.014	0.00008	0.007
<b>KEGG_BASAL_CELL_CARCINOMA</b>	14	0.015	0.00007	0.007
<b>KEGG_FATTY_ACID_METABOLISM</b>	16	0.012	0.0009	0.031
<b>KEGG_STEROID_HORMONE_BIOSYNTHESIS</b>	18	0.014	0.0015	0.031
KEGG_GLYCOPHINGOLIPID_BIOSYNTHESIS_G LOBO_SERIES	5	0.010	0.0008	0.031
KEGG_WNT_SIGNALING_PATHWAY	40	0.017	0.0014	0.031
KEGG_VEGF_SIGNALING_PATHWAY	22	0.014	0.0007	0.031
KEGG_CIRCADIAN_RHYTHM_MAMMAL	5	0.010	0.0012	0.031
KEGG_INSULIN_SIGNALING_PATHWAY	35	0.016	0.0011	0.031
KEGG_OXIDATIVE_PHOSPHORYLATION	32	0.015	0.0022	0.033
KEGG_OTHER_GLYCAN_DEGRADATION	4	0.017	0.0021	0.002
KEGG_ADHERENS_JUNCTION	25	0.014	0.0018	0.030
KEGG_DILATED_CARDIOMYOPATHY	41	0.010	0.0023	0.030
KEGG_SULFUR_METABOLISM	4	0.016	0.0032	0.030
KEGG_DORSO_VENTRAL_AXIS_FORMATION	7	0.014	0.0031	0.007
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS	9	0.015	0.0028	0.007
KEGG_GLYCEROPHOSPHOLIPID_METABOLISM	26	0.012	0.0036	0.031
KEGG_ERBB_SIGNALING_PATHWAY	27	0.014	0.0040	0.031
KEGG_CARDIAC_MUSCLE_CONTRACTION	28	0.010	0.0038	0.031
KEGG_HYPERTROPHIC_CARDIOMYOPATHY_HCM	41	0.017	0.0040	0.031
<b>KEGG_GLIOMA</b>	22	0.014	0.0044	0.031
KEGG_MTOR_SIGNALING_PATHWAY	15	0.010	0.0055	0.031
<b>KEGG_HEDGEHOG_SIGNALING_PATHWAY</b>	12	0.016	0.0058	0.031

**Table 3.** Comparison method result: significant pathways for PDAC prognosis using Global test.

Pathway	Q statistic	Pathway size	p value	q value
<i>KEGG_CIRCADIEN_RHYTHM_MAMMAL</i>	5.271	5	0.0001	0.009
<i>KEGG_PANCREATIC_CANCER</i>	3.215	27	0.0001	0.009
<i>KEGG_CHRONIC_MYELOID_LEUKEMIA</i>	3.012	23	0.0004	0.025
<i>KEGG_COLORECTAL_CANCER</i>	2.692	19	0.0010	0.047

**Table 4.** Comparison method result: significant pathways for PDAC prognosis using Adewale test.

Pathway	W statistic	p value	q value
<i>KEGG_CIRCADIEN_RHYTHM_MAMMAL</i>	34.406	0.0001	0.018
<i>KEGG_PANCREATIC_CANCER</i>	115.407	0.0003	0.018
<i>KEGG_CHRONIC_MYELOID_LEUKEMIA</i>	96.091	0.0002	0.018
<i>KEGG_COLORECTAL_CANCER</i>	73.854	0.0006	0.027
<i>KEGG_MTOR_SIGNALING_PATHWAY</i>	49.852	0.0010	0.029
<i>KEGG_APOPTOSIS</i>	69.954	0.0010	0.029

**Table 5.** Comparison method result: significant pathways for PDAC prognosis using GSEA.

Pathway	Enrichment Score (ES)	Normal ized ES (NES)	p value	q value
<i>KEGG_AUTOIMMUNE_THYROID_DISEASE</i>	-0.478	-1.875	0.002	0.007
<i>KEGG_GLYCOSAMINOGL YCAN_BIOSYNTHESIS_HEPARAN_SULFATE</i>	0.496	1.713	0.013	0.013
<i>KEGG_PRIMARY_IMMUNODEFICIENCY</i>	-0.490	-1.773	0.002	0.017
<i>KEGG_INTESTIL_IMMUNE_NETWORK_FOLICULAR_IGA_PRODUCTION</i>	-0.444	-1.753	0.010	0.020
<i>KEGG_ALLOGRAFT_REJECTION</i>	-0.461	-1.728	0.004	0.024
<i>KEGG_THYROID_CANCER</i>	0.454	1.594	0.018	0.028
<i>KEGG_NOTCH_SIGLING_PATHWAY</i>	0.432	1.591	0.010	0.028
<i>KEGG_HEMATOPOIETIC_CELL_LINEAGE</i>	-0.370	-1.651	0.004	0.043
<i>KEGG_FOCAL_ADHESION</i>	0.306	1.530	0.008	0.044
<i>KEGG_DORSO_VENTRAL_AXIS_FORMATI ON</i>	0.479	1.512	0.044	0.048
<i>KEGG_BLADDER_CANCER</i>	0.410	1.511	0.016	0.048
<i>KEGG_AXON_GUIDANCE</i>	0.315	1.507	0.009	0.050
<i>KEGG_ADHERENS_JUNCTION</i>	0.453	1.923	0.000	0.003
<i>KEGG_ERBB_SIGLING_PATHWAY</i>	0.388	1.714	0.000	0.013

With pathways associated with prognosis, we could also find genes meaningfully related to PDAC prognosis, as well as considering hierarchies of genes and pathways. Table 3 shows genes and pathways significant for the survival phenotype. Using the coefficients of the structural equation, we were able to calculate the  $w_{gene} \times \beta_{path}$  value for each gene. As a result, it was possible to simultaneously consider the effect of the matched gene to the pathway, and the effect size of the pathway to the phenotype. After calculating each coefficient, significance testing was performed, using a permutation method. If the marker was selected based only on a nominal p value, obtained by adapting the entire gene, a type 2 error and a false negative error can be larger. Therefore, we used the False Discovery Rate (FDR) analysis to calculate the q value as a criterion [39]. Interestingly, we also associated the gene *ETS1* with resistance to pancreatic cancer chemotherapy [40]; *ETS-1* also exacerbates poor PDAC prognosis after radiation therapy [41]. Another gene, *HIF1A*, was also noted as a significant indicator of PDAC prognosis in other previous studies [42–44]. Conversely, *GNAI1* was reported as a suppressor of tumor cell migration and invasion that is post-transcriptionally targeted by mir-320a/c/d [45], with the latter being found to confer 5-FU chemo-resistance upon human pancreatic cancer cells [46].

**Table 6.** Significant pathway and gene markers in PDAC prognosis

Pathway	Gene	$w_{\text{gene}} \times \beta_{\text{path}}$	p value	q value
BIOCARTA_G1_PATHWAY	SMAD3	0.032	0.00001	0.004
BIOCARTA_NTH1_PATHWAY	SMAD3	0.032	0.00001	0.004
BIOCARTA_TOB1_PATHWAY	SMAD3	0.032	0.00001	0.004
BIOCARTA_TGFB_PATHWAY	SMAD3	0.032	0.00001	0.004
BIOCARTA_CHEMICAL_PATHWAY	BCL2L1	0.024	0.00003	0.004
BIOCARTA_IL2RB_PATHWAY	BCL2L1	0.024	0.00003	0.004
BIOCARTA_RAS_PATHWAY	BCL2L1	0.024	0.00003	0.004
BIOCARTA_BAD_PATHWAY	BCL2L1	0.024	0.00003	0.004
BIOCARTA_MITOCHONDRIA_PATHWAY	BCL2L1	0.024	0.00003	0.004
BIOCARTA_CTCF_PATHWAY	TGFB1	0.019	0.00005	0.004
BIOCARTA_INFLAM_PATHWAY	TGFB1	0.019	0.00005	0.004
BIOCARTA_ERYTH_PATHWAY	TGFB1	0.019	0.00005	0.004
BIOCARTA_MAPK_PATHWAY	TGFB1	0.019	0.00005	0.004
BIOCARTA_ALK_PATHWAY	TGFB1	0.018	0.00006	0.004
BIOCARTA_G1_PATHWAY	TGFB1	0.018	0.00006	0.004
BIOCARTA_P38MAPK_PATHWAY	TGFB1	0.019	0.00006	0.004
BIOCARTA_TOB1_PATHWAY	TGFB1	0.018	0.00006	0.004
BIOCARTA_NKT_PATHWAY	TGFB1	0.018	0.00006	0.004
BIOCARTA_IL1R_PATHWAY	TGFB1	0.018	0.00006	0.004
BIOCARTA_TGFB_PATHWAY	TGFB1	0.018	0.00006	0.004
BIOCARTA_KERATINOCYTE_PATHWAY	ETS1	0.015	0.00008	0.005
BIOCARTA_ETS_PATHWAY	ETS1	0.015	0.0001	0.006
BIOCARTA_P53HYPOXIA_PATHWAY	HIF1A	0.016	0.00047	0.028
BIOCARTA_HIF_PATHWAY	HIF1A	0.016	0.00047	0.028
BIOCARTA_EPONFKB_PATHWAY	HIF1A	0.016	0.0005	0.028
BIOCARTA_VEGF_PATHWAY	HIF1A	0.015	0.0006	0.033
BIOCARTA_DEATH_PATHWAY	TNFRSF10B	0.018	0.00064	0.033
BIOCARTA_FMLP_PATHWAY	GNAI5	0.015	0.00074	0.037
BIOCARTA_IL1R_PATHWAY	IL1RAP	0.010	0.00095	0.041
BIOCARTA_SET_PATHWAY	GZMA	0.015	0.00100	0.041
BIOCARTA_PTDINS_PATHWAY	PFKP	0.011	0.00110	0.041
BIOCARTA_EXTRINSIC_PATHWAY	TFPI	0.013	0.00115	0.041
BIOCARTA_AML_PATHWAY	TFPI	0.013	0.00116	0.041
BIOCARTA_PAR1_PATHWAY	GNAI1	0.017	0.00118	0.041
BIOCARTA_EDG1_PATHWAY	GNAI1	0.017	0.00119	0.041
BIOCARTA_GPCR_PATHWAY	GNAI1	0.017	0.00119	0.041
BIOCARTA_SPPA_PATHWAY	GNAI1	0.017	0.00122	0.041
BIOCARTA_BIOPEPTIDES_PATHWAY	GNAI1	0.017	0.00122	0.041
BIOCARTA_CXCR4_PATHWAY	GNAI1	0.017	0.00122	0.041
BIOCARTA_MPR_PATHWAY	GNAI1	0.017	0.00122	0.041
BIOCARTA_PPARA_PATHWAY	ACOX1	0.015	0.00122	0.041
BIOCARTA_GSK3_PATHWAY	GNAI1	0.017	0.00123	0.041
BIOCARTA_VEGF_PATHWAY	VEGFA	0.010	0.00146	0.047
BIOCARTA_NO1_PATHWAY	VEGFA	0.010	0.00147	0.047
KEGG_CELL_CYCLE	SMAD3	0.023	0.00010	0.047
KEGG_WNT_SIGNALING_PATHWAY	SMAD3	0.023	0.00010	0.047
KEGG_TGF_BETA_SIGNALING_PATHWAY	SMAD3	0.023	0.00010	0.047



# Chapter 5.

## Discussions

Among pathway analysis methods, few have been developed only for survival times. Thus, there is a need for a way to quantitatively determine how much the pathway affects the survival phenotype and identify a relative way of ranking pathways. To this end, HisCoM-PAGE uses structural equations to model real biological phenomena, so it can estimate not only the value of statistics of a pathway, but also meaning of pathway statistics. Because the  $\beta$  parameter is a hazard ratio of pathway for survival phenotype, the effect of the pathway on the disease can be analyzed by considering both the magnitude and the sign of the coefficient representing the pathway.

Consequently, we performed survival analysis using data from PDAC patients with poor prognoses, first by identifying prognosis-related pathways, and then by further analysis to find specific genes associated with survival times. Thus, finding pathways related to prognosis, through HisCoM-

PAGE, looks not only at association with survival times, but also how genes behave within a biological structure. Among the important pathways found in this study, many could play important roles in studying the prognosis of PDAC, including the  $TGF-\beta$  and Hedgehog signaling pathways, and we also validated the genes with important roles in prognosis-related pathways. Furthermore, we also found significant genes, in addition to the genes mentioned in the previous results, such as *SMAD3*, *BCL2*, and *TGF- $\beta$  1* [47–51].

Recently, many studies have been actively conducted to examine PDAC prognosis using RNAseq data [52,53], which is easily processed by our HisCoM-PAGE, to find prognosis-related pathways. For example, RNA-seq data could be used to define new latent variables by applying many clustering methods into our HisCoM-PAGE model. Such an application could overcome the limitations of other pathway methods that rely solely on pathway databases.

Beyond looking only at associations for survival phenotype, we can also use the HisCoM-PAGE model to construct other types of predictive models for prognosis. We could study the design of a prognostic prediction model using the latent variable pathway as a marker, as well as the genetic marker [54]. In this case, unlike building predictive models using

only genetic markers, we can add interpretability because the designed predictive model considers the contribution of genetic markers in a pathway manner. Also, it would be possible to study the relationship between genes and pathways, beyond the linear relationship, using the kernel generalized structured component analysis (GSCA) method.

In summary, in this study, we proposed a new pathway analysis method, Hierarchical Structured Component Analysis of Pathway Analysis for Gene Expression (HisCoM-PAGE), to identify disease prognosis-related pathways. By assessment, using simulated data and PDAC microarray data, HisCoM-PAGE performed better than other pathway-based methods. Moreover, HisCoM-PAGE could also find more interpretable and meaningful pathways and prognostic genetic markers. Thus, we believe that the HisCoM-PAGE can easily be extended to other types of gene expression data, such as RNA sequencing data, and that such analyses could be quite valuable in the modern-day era of precision medicine.

# Bibliography

- 1 LOCKHART, David J.; WINZELER, Elizabeth A. Genomics, gene expression and DNA arrays. *Nature*, 2000, 405.6788: 827.
- 2 CASAMASSIMI, Amelia, et al. Transcriptome profiling in human diseases: new advances and perspectives. *International journal of molecular sciences*, 2017, 18.8: 1652.
- 3 BYRON, Sara A., et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 2016, 17.5: 257.
- 4 SOTIRIOU, Christos; PUSZTAI, Lajos. Gene-expression signatures in breast cancer. *New England Journal of Medicine*, 2009, 360.8: 790–800.
- 5 KHATRI, Purvesh; SIROTA, Marina; BUTTE, Atul J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 2012, 8.2: e1002375.
- 6 MACIEJEWSKI, Henryk. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics*, 2013, 15.4: 504–518.
- 7 LIU, Qi, et al. Comparative evaluation of gene-set analysis methods. *BMC bioinformatics*, 2007, 8.1: 431.
- 8 KIM, Bohyeon; HA, Il Do; LEE, Donghwan. Analysis of multi-center bladder cancer survival data using variable-selection method of multi-level frailty models. *Journal of the Korean Data and Information Science Society*, 2016, 27.2: 499–510.
- 9 RYU, Ji Kon. The Early Detection of Pancreatic Cancer: Whom and How?. *The Korean Journal of Pancreas and Biliary Tract*, 2015, 20.4: 198–203.

- 10 SUBRAMANIAN, Aravind, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 2005, 102.43: 15545–15550.
- 11 GOEMAN, Jelle J., et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 2004, 20.1: 93–99.
- 12 GOEMAN, Jelle J., et al. Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 2005, 21.9: 1950–1957.
- 13 ADEWALE, Adeniyi J., et al. Pathway analysis of microarray data via regression. *Journal of Computational Biology*, 2008, 15.3: 269–277.
- 14 HÄNZELMANN, Sonja; CASTELO, Robert; GUINNEY, Justin. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, 2013, 14.1: 7.
- 15 LEE, Sungyoung, et al. Pathway-based approach using hierarchical components of collapsed rare variants. *Bioinformatics*, 2016, 32.17: i586–i594.
- 16 HWANG, Heungsun; TAKANE, Yoshio. Generalized structured component analysis. *Psychometrika*, 2004, 69.1: 81–99.
- 17 KIM, Yongkang, et al. Hierarchical structural component modeling of microRNA-mRNA integration analysis. *BMC bioinformatics*, 2018, 19.4: 75.
- 18 CHOI, Sungkyoung, et al. HisCoM-GGI: Hierarchical structural component analysis of gene-gene interactions. *Journal of bioinformatics and computational biology*, 2018, 1840026–1840026.
- 19 LEE, Seungyeoun; KIM, Jinheum; LEE, Sunho. A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC bioinformatics*, 2011,

- 12.1: 377.
- 20 WAN, Fei. Simulating survival data with predefined censoring rates for proportional hazards models. *Statistics in medicine*, 2017, 36.5: 838–854.
  - 21 DE LEEUW, Christiaan A., et al. The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 2016, 17.6: 353.
  - 22 KAO, Patrick YP, et al. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochimica et Biophysica Acta (BBA)–General Subjects*, 2017, 1861.2: 335–353.
  - 23 IRIZARRY, Rafael A., et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003, 4.2: 249–264.
  - 24 MARCZYK, Michal, et al. Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition. *BMC bioinformatics*, 2013, 14.1: 101.
  - 25 MCCLINTICK, Jeanette N.; EDENBERG, Howard J. Effects of filtering by Present call on analysis of microarray experiments. *BMC bioinformatics*, 2006, 7.1: 49.
  - 26 CALZA, Stefano, et al. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. *Nucleic acids research*, 2007, 35.16: e102.
  - 27 OGATA, Hiroyuki, et al. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 1999, 27.1: 29–34.
  - 28 NISHIMURA, Darryl. BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, 2001, 2.3: 117–120.
  - 29 SHEN, Wei, et al. TGF- $\beta$  in pancreatic cancer initiation and progression: two sides of the same coin. *Cell & bioscience*, 2017, 7.1: 39.

- 30 TRUTY, Mark J.; URRUTIA, Raul. Basics of TGF- $\beta$  and pancreatic cancer. *Pancreatology*, 2007, 7.5-6: 423-435.
- 31 DERYNCK, Rik; AKHURST, Rosemary J.; BALMAIN, Allan. TGF- $\beta$  signaling in tumor suppression and cancer progression. *Nature genetics*, 2001, 29.2: 117.
- 32 FRIESS, Helmut, et al. Enhanced expression of transforming growth factor  $\beta$  isoforms in pancreatic cancer correlates with decreased survival. *Gastroenterology*, 1993, 105.6: 1846-1856.
- 33 WAKEFIELD, Lalage M.; ROBERTS, Anita B. TGF- $\beta$  signaling: positive and negative effects on tumorigenesis. *Current opinion in genetics & development*, 2002, 12.1: 22-29.
- 34 VILLANUEVA, Alberto, et al. Disruption of the antiproliferative TGF- $\beta$  signaling pathways in human pancreatic cancer cells. *Oncogene*, 1998, 17.15: 1969.
- 35 JAVLE, Milind, et al. Biomarkers of TGF- $\beta$  signaling pathway and prognosis of pancreatic cancer. *PloS one*, 2014, 9.1: e85942.
- 36 KUNDU, Juthika, et al. Tob1 induces apoptosis and inhibits proliferation, migration and invasion of gastric cancer cells by activating Smad4 and inhibiting  $\beta$ -catenin signaling. *International journal of oncology*, 2012, 41.3: 839-848.
- 37 WANG, Jin; SEN, Subrata. MicroRNA functional network in pancreatic cancer: from biology to biomarkers of disease. *Journal of biosciences*, 2011, 36.3: 481-491.
- 38 LU, Yuan, et al. Genes targeted by the Hedgehog-signaling pathway can be regulated by Estrogen related receptor  $\beta$ . *BMC molecular biology*, 2015, 16.1: 19.
- 39 BENJAMINI, Yoav; HOCHBERG, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple

- testing. *Journal of the Royal statistical society: series B (Methodological)*, 1995, 57.1: 289–300.
- 40 XI, Ting; ZHANG, Guizhi. Integrated analysis of tumor differentiation genes in pancreatic adenocarcinoma. *PloS one*, 2018, 13.3: e0193427.
  - 41 TOMIHARA, Hideo, et al. MicroRNA–181b–5p, ETS1, and the c–Met pathway exacerbate the prognosis of pancreatic ductal adenocarcinoma after radiation therapy. *Cancer science*, 2017, 108.3: 398–407.
  - 42 SHIBAJI, Takamune, et al. Prognostic significance of HIF–1 alpha overexpression in human pancreatic cancer. *Anticancer research*, 2003, 23.6C: 4721–4727.
  - 43 HOFFMANN, Andreas–Claudius, et al. High expression of HIF1a is a predictor of clinical outcome in patients with pancreatic ductal adenocarcinomas and correlated to PDGFA, VEGF, and bFGF. *Neoplasia*, 2008, 10.7: 674–679.
  - 44 SUN, Hong–Cheng, et al. Expression of hypoxia–inducible factor–1 alpha and associated proteins in pancreatic ductal adenocarcinoma and their impact on prognosis. *International journal of oncology*, 2007, 30.6: 1359–1367.
  - 45 YAO, Jian, et al. GNAI1 suppresses tumor cell migration and invasion and is post–transcriptionally regulated by Mir–320a/c/d in hepatocellular carcinoma. *Cancer biology & medicine*, 2012, 9.4: 234.
  - 46 WANG, Weibin, et al. MicroRNA–320a promotes 5–FU resistance in human pancreatic cancer cells. *Scientific reports*, 2016, 6: 27641.
  - 47 YAMAZAKI, Ken, et al. Upregulated SMAD3 promotes epithelial–mesenchymal transition and predicts poor prognosis



- in pancreatic ductal adenocarcinoma. *Laboratory investigation*, 2014, 94.6: 683.
- 48 UNGEFROREN, Hendrik, et al. Differential roles of Smad2 and Smad3 in the regulation of TGF- $\beta$ 1 mediated growth inhibition and cell migration in pancreatic ductal adenocarcinoma cells: control by Rac1. *Molecular cancer*, 2011, 10.1: 67.
  - 49 SONG, Shanshan, et al. Expression of Beclin 1 and Bcl-2 in pancreatic neoplasms and its effect on pancreatic ductal adenocarcinoma prognosis. *Oncology letters*, 2017, 14.6: 7849–7861.
  - 50 RAY, Katrina. Pancreatic cancer: new insights into PDAC growth promotion via a BAG3-mediated paracrine loop. *Nature Reviews Gastroenterology & Hepatology*, 2015, 12.12: 669.
  - 51 ZHAO, J., et al. Clinical and prognostic significance of serum transforming growth factor- $\beta$ 1 levels in patients with pancreatic ductal adenocarcinoma. *Brazilian Journal of Medical and Biological Research*, 2016, 49.8.
  - 52 JANKY, Rekin' s, et al. Prognostic relevance of molecular subtypes and master regulators in pancreatic ductal adenocarcinoma. *BMC cancer*, 2016, 16.1: 632.
  - 53 RAPHAEL, Benjamin J., et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, 2017, 32.2: 185–203. e13.
  - 54 KIM, Sungtae, et al. Drug response prediction model using a hierarchical structural component modeling method. *BMC bioinformatics*, 2018, 19.9: 117.

## 초 록

암에 상관관계가 있는 생물학적 기작 곧, 패스웨이를 찾아내기 위한 여러 가지 분석이 있었지만 유전자 발현 데이터를 기반으로 한 분석들의 대부분은 단일 패스웨이 분석에 기초하고 있었다. 이러한 분석 방법의 경우, 패스웨이들 간의 상관 관계를 고려하지 않았다. 본 논문에서는 유전자와 그 상위 단계라고 할 수 있는 패스웨이의 생물학적인 위계 구조를 반영하는 HisCoM-PAGE: 계층적 구조 모형을 이용한 유전자 발현 데이터의 패스웨이 분석 모델을 제안한다. 특히, HisCoM-PAGE는 생존자료 표현 형에 초점을 맞추고 예후에 상관관계를 가지는 통계적으로 유의한 패스웨이를 찾아내는 것에 중점을 두었다. 실제 데이터에 대한 적용으로는 췌장암 데이터를 이용하였는데, 이는 췌장암이 여러 암 중 중에서도 예후가 좋지 못한 질병으로, 예후에 대한 연구가 중요하기 때문이다. HisCoM-PAGE 방법을 실제 췌장암 유전자 발현 데이터에 적용하였을 때, HisCoM-PAGE 방법이 췌장암 예후와 관련된 패스웨이를 효과적으로 찾아낼 수 있다는 것을 확인하였다. 또한, 제시한 방법론의 통계적인 검정력을 확인하기 위해서 기존에 패스웨이 방법론으로 제안된

Gene Set Enrichment Analysis(GSEA), Global Test(GT), Adewale Test 와 같은 다른 패스웨이 방법론과 비교하여 시뮬레이션 연구를 진행하였다. 타 방법론과의 비교를 통해서 HisCoM-PAGE가 질환과의 상관 관계를 가지는 통계적으로 유의한 패스웨이를 찾아내는데 높은 검정력을 가지는 것을 확인하였다.

**주요어:** 패스웨이 분석, 생존 분석, 계층적 구조 모형

**학 번:** 2017 -29730